DOCUMENT RESUME

ED 339 744 TM 017 676

AUTHOR Zin, Than Than; Williams, John
TITLE Searching for Better Scoring of Multiple-Choice Tests: Proper Treatment of Misinformation, Guessing and Partial Knowledge.
PUB DATE 91
NOTE 15p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Educational Research; Estimation (Mathematics); *Guessing (Tests); Literature Reviews; Models; *Multiple Choice Tests; *Scoring; Testing Problems; Test Items; Test Theory; *Test Wiseness
IDENTIFIERS Ability Estimates; *Finite State Score Theory; *Partial Knowledge (Tests)

ABSTRACT

Brief explanations are presented of some of the different methods used to score multiple-choice tests; and some studies of partial information, guessing strategies, and test-taking behaviors are reviewed. Studies are grouped in three categories of effort to improve scoring: (1) those that require extra effort from the examinee to answer multiple-choice questions compared to the conventional mode of selecting the best or correct option among two or more choices; (2) those that manipulate test items or instructions; and (3) those that analyze original responses in special ways. The alternative scoring models of these studies have not provided a comprehensive picture of response behaviors because the manner in which they have been set up is fragmentary, and their assumptions are specific for the response mode for which the model was designed. A more generalized modeling approach is apparent in the finite state score theory of performance on multiple-choice tests proposed by M. A. Garcia-Perez (1987). Finite state score theory parsimoniously incorporates assumptions appropriate to the particular mode in which the test is administered, and it yields ability estimates on a single metric that remains the same regardless of the mode of test administration or assumptions concerning examinee behavior. The theory awaits empirical testing, but seems to be adaptable to any assumptions about item characteristics. Consequently, it appears to be a promising trend in measurement. A 31-item list of references is included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

JOHN WILLIAMS

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED339744

SEARCHING FOR BETTER SCORING OF MULTIPLE-CHOICE TESTS: PROPER TREATMENT OF MISINFORMATION, GUESSING AND PARTIAL KNOWLEDGE

Than Than Zin and John Williams
Virginia Tech

TM017676

BEST COPY AVAILABLE

Searching for better scoring of multiple-choice tests:
Proper treatment of misinformation, guessing and partial knowledge

Than Than Zin and John Williams
Virginia Tech

Most measurement experts and testing organizations prefer to use the multiple-choice format for standardized cognitive tests in education and other areas because such tests provide broad content sampling, high reliability, ease of administration and scoring, and usefulness in testing varied content (Haladyna & Downing, 1989). It is also one of the most common and favored test item types for classroom teachers (Marso,1985). Contrary to popular belief, multiple-choice tests are not limited to testing recognition and rote memorization but may also measure higher level thinking (Haladyna & Downing, 1989).

However, the scoring of multiple-choice tests has been problematical. Dichotomous scoring of multiple-choice items causes irregularities in the interpretation of test scores because the response to a multiple choice item may come from different information levels, namely, misinformation, total ignorance, partial information, and complete information. Guessing strategies and willingness to guess when an examinee does not know the answer also contribute to an item response. Dichotomous scoring has also been criticized because it usually fails to provide a direct estimate of the amount of knowledge examinees have. Indeed, most multiple-choice tests provide only ranking information.

For over fifty years, the substantial weight of evidence has been that examinees are able to make use of partial information when responding to multiple-choice items (Hutchinson, 1985). Jaradat & Tollefson (1988) suggested that different

TM017676

scoring methods (those which account for partial knowledge, and those which do not) differ substantially in the proportions assigned to different grade categories for criterion referenced grading decisions. To account for the possibility of partial knowledge or misinformation, we should assign different item scores to the various incorrect responses.

The examinees also differ in their willingness to guess or equivalently, willingness to omit. The finding of Noble et al., (1986) showed that the rate of increase of item omissions, as well as the omit rate itself, were both significantly different for students from various ethnic groups on criterion referenced tests. Because of this, "number right" scores measure, not only the examinees' ability, but also the examinee's willingness to guess or omit the item. For these reasons it is essential to treat misinformation, partial information, and guessing properly to provide test scores that yield better classification and estimation of student ability levels.

In the psychometric literature, there have been many attempts to overcome these weaknesses and formulate procedures for extracting information from the examinee selection of incorrect responses to multiple-choice items. These studies have been done with different kinds of scoring methods under different modes of test administration with different types of cognitive tests to understand guessing, partial knowledge, and misinformation.

Although the studies in this area have a long history, studies in the 1980's will be the primary focus of this paper. However, some older studies are described to provide different perspectives. Another purpose of this paper is to introduce one of the recently developed theories in this area which tries to control all these factors in an interesting

way. Thorough review of the studies in this area can be found in Frary (1989) and Hutchinson (1991). In this paper, a brief explanation of some of the different scoring methods , some studies on partial information, guessing strategies, test taking behaviors, and findings of them will be presented. Although the categories are not absolutely exclusive, I would like to present these studies in three groups, according to the primary nature of their strategies. These are 1. those that require extra effort from examinees; 2. those that manipulate test items or instructions; and 3. those that simply analyze conventional responses in special ways.

1. those that require extra effort from examinees

Studies in the first category require extra effort from an examinee to answer multiple choice items compared to the conventional response mode, i.e., selecting the option which an examinee believes is the best or correct answer among two or more choices.

One of the studies in this category provided a scoring method, namely confidence rating of all options. De Finetti (1965) proposed the scoring method in which an examinee was required to express partial knowledge as the estimated probability of correctness of each option. These estimates are called subjective probabilities. Recent studies of this method can be seen in Rippey and Voytovich (1985). Although this method provides more information than "number right" scores, determining how to interpret them is complicated. Similar to this method is confidence rating of a single selected option within each item. This method asks an examinee to select the perceived option as most likely to be correct and, at the same time, express confidence in that

choice. Detailed discussion on this method can be found in Leclercq (1984). He concluded that sophisticated tools are essential to get quality data.

Some studies in this category used answer until correct and other similar scoring methods. These methods ask examinees to continue selecting options until the correct answer is chosen by using special answer sheets or media. This method allows the study of psychometric properties of items and total scores from a test. For example, Wilcox & Wilcox (1988) used this method to study decision making processes for multiple-choice items. However, this method has not been practical because of scoring labor requirements, high administration costs due to the special media, and the need for examinee experience.

Some studies used subset selection scoring methods. In one version of this method, examinees are instructed to identify the subset of options that include all the options that they believe are distractors. In another, they are asked to include all options they believe necessary to ensure inclusion of the correct option(s). Studies based on this method were first reported by Coombs (1952) and Dressel & Schmid (1953). Echternacht (1972) evaluated this method in his review and concluded that early studies in this method produced uniform improvement in internal consistency reliability but no improvement in validity compared to "number right" or formula scored versions of the same test. The same conclusion can be found in the simulation study of Frary (1982). A recent study on this method can be found in Jaradat and Sawaged (1986). In contrast to most research on this method, this study reported that scores computed from the Dressel and Schmid (1953) method had better reliability and validity and did not tend to favor high

risk takers compared to number right and formula scores.

Some of the studies on partial knowledge, guessing, test taking behaviors, in this category were based on student reluctant responses. In one of these studies, a examiner insisted that students give responses which were originally omitted Bliss (1980). A similar method to this strategy ,answer changing, can be found in the review of Benjamin, Cavell, & Shallenberger (1984). The findings of these studies and earlier studies consistently suggested that examinees were able to use partial knowledge and productive guessing strategies.

2. those that manipulate test items or instructions

The second category in this area involves certain kinds of manipulation of the test or test items or of the instructions to the test. One of these studies Webb, Herman, Cabello, (1986) used matched pairs of parallel items in a test of pronoun usage to diagnose students' error. Contrary to previous findings they found that young students (upper elementary) were not able to use partial knowledge. Some studies used multiple true-false items, that is an item consisted of a stem, together with more than one statement, from which an examinee had to classify it true or false, with regard to the stem. The number of options varied from one item to another. Hsu, Moss, Khampalikit (1984) used this type of item and studied six different scoring methods. They found the method which accounted for partial knowledge yielded more reliable and valid scores.

Traub and Hambelton (1972) studied formula scores, based on the same test with three different instructions: The first one was rewarding instructions for omitted items,

the second was threatening a penalty for wrong, and the third was a conventional number right instruction. They found that rewarding instruction was more effective in encouraging omissive behaviors than penalty instructions. The rewarding instructions yielded higher reliability test scores than other two. Although the instructions had no significant effect on criterion validity, rewarding instructions yielded higher validity scores than other two. Plake, Wise, Harvey, (1988) studied the test taking behaviors of examinees under different instructions. They concluded that students (undergraduate college students) not only comprehend instructions but can calculate how much advantage they can gain from different item-response strategies.

To study guessing behaviors, some studies in this category used nonsense items; That is the item which had no correct answer among the available options or the question stem itself was meaningless. They used the proportion of nonsense items which subjects attempted as a measure of an examinee's willingness to guess to interpret risk taking behavior on examination Slakter (1969). Hutchinson (1985) studied the relationship between the responses to nonsense items and those of genuine items based on the theory. He found fairly good agreement between observed and predicted proportion of the nonsense items attempted from the proportion of genuine items answered correctly, incorrectly, and omitted .

3. those that simply analyze original responses in special ways.

The studies in the third category do not manipulate either items or instructions. They do not extract more information from examinees. These techniques use

the responses from an ordinary multiple-choice test administered under normal instructions. The earliest technique was based on the assumption that, if an examinee did not know the answer, they would either omit the item or guess at random. This thinking led to a formula to "correct for guessing". A proportion of the number of wrong answers (supposedly achieved correctly by guessing) was subtracted from the "number right". Examinees who had partial knowledge about an item did not respond at random nor did examinees with misinformation about that item. Wrong options may not be equally attractive to the examinees in these cases (Lord & Novick, 1968).

Another method in this category is response weighting (option weighting). This method involves giving different scores (weights) to all of the options. Two different kinds of weights are used in this method. One consists of weights provided by judges in terms of the degree of correctness of an option and the other is empirical weighting, based on the responses of the examinees themselves. Smith (1987) has done an empirical study on this method with a vocabulary test. He determined response weights a priori, based on a theory, assigned different scores for different wrong answers and estimated examinee ability. The results indicated that reliability and concurrent validity of response weighting scores were significantly higher than 'number-right' scores on the same test.

Item response theory also treats guessing by using a three parameter logistic model. The study of Item response theory for partial knowledge can be found in Masters (1982), which required that responses be ranked in terms of correctness.

<u>The trend toward finite state score theory</u>

Some of the above mentioned methods, such as confidence weighting and response weighting, showed consistent increases in internal consistency reliability at the expense of longer testing time or undue scoring and administration complexity. Some of the methods, such as answer until correct, provide immediate feedback that may enhance the learning process. Other methods, such as manipulating test instruction, provided some information to understand the test taking behavior or guessing behaviors of examinees. However, these techniques have not provided more meaningful scores which directly estimate student ability (except the item response theory method) with better reliability and/or validity than the conventional number right score.

For this reason Frary (1989) commented that none of these attempts were uniformly desirable or efficacious regarding examinee acceptance, resource usage and psychometric properties of scores. These alternative scoring models fail to provide a comprehensive picture of response behaviors because the way in which these alternative models have been set up is fragmentary and the assumptions made in each of them are specific for the particular response mode in which the model was designed (Garcia-Perez 1991).

Recent development of strategies in this area involve statistical modeling based on a theory, without manipulating test items or test instructions nor collecting more information from students. García-Pérez (1987) proposed a finite state theory of performance on multiple-choice tests using a more generalized modeling approach, based on earlier work by Hutchinson (1982). Compared to conventional scoring schemes, scores derived from finite state models have two major advantages.

First, finite state theory parsimoniously incorporates assumptions appropriate to the particular mode in which the test is administered. These assumptions concern subjects' item response behavior when responding to a multiple-choice item, rather than merely accounting for their final responses. Secondly, the theory yields ability estimates on a single metric which remains the same regardless of the mode of test administration or assumptions concerning examinee behavior.

Finite state theory accounts for partial knowledge by shifting attention from the item to the option. The term "statement" refers to any of the sentences that result from adding to an item stem any of its alternative options. The theory assumes that examinees attempt to classify each option within each item as true or false, instead of responding holistically to the item. It also assumes that it is their knowledge which determines whether or not they select the correct option on the answer sheet. Finite state theory recognizes the existence of all-or-none knowledge about corresponding options.

Finite state scoring models derived from the theory are adaptable to any assumptions about item characteristics and examinee response strategies. Use of these models yields finite state scores, which are nonlinear transformations of the proportions of conventional multiple-choice responses that are correct, incorrect, and embedded omissions. The latter are defined as leaving the item unanswered in the middle of the test. Trailing omissions, the omissions that occur at the end of test, would indicate speededness.

Moreover, these estimates of finite state scores are ratio measures that are directly interpretable as the amounts of knowledge or ability that examinees have (Garcia Perez

1987). Specifically, a finite state score estimates the percentage of statements from a universe of items that the examinee can classify accurately as correct or incorrect without seeing the options together in the items. These estimates are equivalent across differing assumptions about item type, administration conditions and examinee behavior.

For example, García-Pérez (1987) showed that the ability estimates of finite state scoring methods derived for different modes of test administration (e.g., Coombs response scoring, formula scoring) were equivalent. The simulation study by García-Pérez and Frary (1989) was highly supportive of the use of finite state scores.

Although the theory allows us to make alternative assumptions about item characteristics and examinee response strategies, it is necessary to employ a procedure to check the validity of these assumptions. The procedure was applied most recently by García-Pérez and Frary (1991) with respect to items containing the option "none of the above". It calls for additional empirical studies using this theory to gain a view of how the theory works in practice.

While we await continued empirical testing of the finite state score model, measurement specialists may continue to experiment with score correction procedures that 1. require extra effort from examinees; 2. manipulate test items or instructions; and 3. simply analyze conventional responses in special ways. Of the many methods developed in the last fifty years or so of research, finite state score theory seems to be a promising trend in this area of measurement study.

References

Benjamin, L.T., Cavell, T.A., & Shallenberger, W.R. (1984). Staying with initial answer on objective tests: Is it a myth? Teaching of Psychology , 11, 133-141.

Bliss, L.B. (1980). A test of Lord's assumption regarding examinee guessing behavior on multiple choice tests using elementary school students. Journal of Educational Measurement , 17, 147-153.

Coombs, C.H. (1953). On the use of objective examinations. Educational and Psychological Measurement, 13, 309-310.

Dressel, P.L., & Schmid, J. (1953). Some modifications of the multiple choice item. Educational and Psychological Measurement , 13, 574-595.

De Finetti, B., (1965). Methods for discriminating levels of partial knowledge concerning a test item. British Journal of Mathematical & Statistical Psychology , 18, 87-123.

Echternacht, G.J., (1972). The use of confidence testing in objective testing. Review of Educational Research , 42, 217-236.

Frary, R.B. (1989).Partial credit scoring methods for multiple choice tests. Applied Measurement in Education, 2, 79-96.

Frary, R.B. (1980). The effect of misinformation, partial information, and guessing on expected multiple-choice test item scores. Applied Psychological Measurement , 4, 79-90.

Garcia-Perez, M.A. & Frary, R.B. (1991). Testing finite models of performance in multiple-choice tests using items with "none of the above" as an option. In J.-C. Falmagne & J.-P. Doignon (Eds.), Mathematical Psychology: Current developments. New York: Springer-Verlag.

Garcia-Perez, M.A. (1990). A comparison of two models of performance in objective tests: Finite state versus continuous distributions. British Journal of Mathematical and Statistical Psychology, 43, 73-91.

Garcia-Perez, M.A. & Frary, R.B. (1989). Psychometric properties of finite-state scores versus number-correct and formula scores: A simulation study. Applied Psychological Measurement, 13, 403-417.

Garcia-Perez, M.A. (1987). A finite state theory of performance in multiple choice tests. In E.E.Roskam & R.Suck (Eds.), Progress in mathematical psychology-I (pp. 455-462). Amsterdam: Elsevier.

Haladyna, T.M. & Downing, S.M. (1989). A taxonomy of multiple-choice item-writing rules. Applied Measurement in Education, 2, 37-50.

Hsu, T.C., Moss, P.A., & Khampalikit, C.(1984). The merits of multiple-answer item as evaluated by using six scoring formulas. Journal of Experimental Education , 52(3), 152-158.

Hutchinson, T.P. (1991). Ability, Partial Information, Guessing: Statistical Modelling Applied to Multiple-Choice test.(PP. 81-138). Adelaide, South Australia: Rumsby.

Hutchinson, T.P. (1985 July). Predicting performance in variants of multiple choice test. Paper presented at the annual meeting of the Psychometric Society and the Classification Societies. Fourth Cambridge, England.

Hutchinson, T.P. (1982). Some theories of performance in multiple choice tests and their implications for variants of task. British Journal of Mathematical and Statistical Psychology , 35, 71-89.

Jaradat, D. & Sawaged (1986). The sub set selection techniques for multiple choice tests.An empirical inquiry. Journal of Educational Measurement, 23, 369-376.

Jaradat, D. & Tollefson, N. (1988).The impact of alternative scoring procedures for multiple choice items on test reliability validity and grading. Educational and Psychological Measurement, 48(3), 627-635.

Leclercq, D. (1984). Confidence marking: its use in testing. In B.H.Choppin & T.N. Postlethwaite (Eds.), Evaluation in Education. An international Review series: Vol.6 (2). (P.P 163-287). Oxford: Pergomon Press.

Lord, F.M. & Novick, M.R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Marso, R.N. (1985 October) Testing practice and test item preferences of classroom teachers. Paper presented at the annual meeting of the Mid Western Educational Research Association , Chicago, IL.

Master, G.N. (1982). A rasch model for partial credit scoring. Psychometrika , 47, 149-174.

Nobel, C. et al, (1986 April). Omit rates on Criterion-Referenced Tests for Different Ethnic groups: Implication for large scale assessment. Paper presented at the 67th annual meeting of the American Educational Research Association San Francisco, CA.

Plake, B.S., Wise, S.L. & Harvey, A.L. (1988). Test taking behavior under formula and number-right scoring conditions. Bulletin of the Psychonomic Society , 26, 316-318.

Rippey, R.M. & Voytovich, A.E. (1985). Anomalous responses on confidence scored tests. Evaluation and the Health Professions, 8, 109-119.

Slakter, M.J. (1969). Generality of risk taking on objective examinations. Educational and Psychological Measurement , 29, 115-128.

Smith, R.M. (1987). Assessing partial knowledge in vocabulary. Journal of Educational Measurement , 24, 217-231.

Traut, R.E. & Hambleton, R.K. (1972). The effort of scoring instructions and degree of speededness on the validity and reliability of multiple choice tests. Educational and Psychological Measurement , 32, 737-758.

Webb, N.M., Herman, J.L. & Cabello, B.(1986). Diagnosing students errors from their response selections in language arts. Journal of Educational Measurement , 23, 163-170.

Wilcox, R.R & Wilcox, K.T. (1988). Models of decision making processes for multiple choice test items: An analysis of spatial ability. Journal of Educational Measurement , 25(2), 125-136.